Model-Agnostic Private Learning

Raef Bassily1, Om Thakkar2, Abhradeep Thakurta3

1The Ohio State University, 2Boston University, 3University of California-Santa Cruz

INTRODUCTION

This work provides a framework, using black-box transformations of non-private learners, for obtaining: 1) Privacy-preserving predictions, and 2) A private classifier from private predictions.

Preliminaries

If \(\mathcal{B},\mathcal{D}\)-PAC learner: Alg. \(\mathcal{A}_0\). \(\mathcal{B}\in\{+,-\}\), (agnostic) PAC learner for hypothesis class \(\mathcal{H}\) given input dataset \(D\rightarrow\mathcal{D}'\), e.g. \(1\rightarrow\mathcal{D}\) outputs a hypothesis \(\mathcal{A}_0\in e\mathcal{D}\) with where \(err(D,\mathcal{D}')\) denotes the misclassification rate of \(\mathcal{A}_0\) on \(D\), and \(\gamma\) is the complexity of \(\mathcal{A}_0\).

We call it the realizable case if \(\gamma = 0\), else we call it the agnostic case.

\(\mathcal{B}\), \(\mathcal{D}\)-Differentially Private (DP) learner: A randomized algorithm \(\mathcal{A}_0\). \(\mathcal{B}\in\{+,-\}\) or \(\mathcal{B}\) for different hypotheses \(D\rightarrow\mathcal{D}\). \(\mathcal{D}\in\mathcal{D}'\), \(\mathcal{D}'\rightarrow\mathcal{D}'\), \(\gamma\) is the complexity implied by our construction.

Advantages of the framework:

- Makes black-box use of non-private learners
- Conservatively uses privacy budget to label lots of public features \([PAE+\,17]\).
- Allows knowledge transfer using public features and output labels to train a private classifier.
- Results in transferable utility guarantees for all queries.

Non-private learner accuracy = Private learner accuracy + \(\delta\).

Theorem: \(\delta\) is a black-box transformation, but didn’t provide formal accuracy guarantees. However, the experiments in [WBIH19] indirectly corroborate our intuition and theory as stability.

Private Algorithm for Classification Queries

For each query, if margin is sufficiently large, the output is the majority vote.

Privacy: To ensure \(\epsilon,\delta\) DP, private alg. \(\mathcal{A}_0\) takes after \(\epsilon/3\) margin optimistic predictions. The privacy budget is only consumed by unlabelable predictions.

Generic Transformation of Misclassification Rate:

\[\text{Misclassification rate of } \mathcal{A}_0 \text{ is } \frac{3}{2}\times\epsilon \text{ for a specific setting of } k.\]

\[\text{Proof idea: In each classifier } \mathcal{A}_0 \text{, misclassification rate is } \epsilon, \text{ then by a counting argument, at least } 2\epsilon \text{ classifiers will agree on the correct label, except for } \epsilon \text{ time queries.}\]

\[\text{Our algorithm can answer } 1/\epsilon \text{ more queries than prior approaches based on the "composition theorem" of differential privacy.}\]

Building Blocks

Goal: A model for binary classification, e.g., predicting Parkinson’s disease.

1. Private Algorithm for Classification Queries

2. From Private Predictions to a Private Classifier

REFERENCES